
A TERABIT MULTISERVICE SWITCH

THE CYCLONE SWITCH ARCHITECTURE ENABLES A SCALABLE SWITCHING PLATFORM FROM MULTIPLE GBITS TO MULTIPLE TBITS PER SECOND IN FIVE CUSTOM 0.18-MICRON CMOS ICs. A WIRE-SPEED SCHEDULING CAPABILITY SUPPORTS EIGHT QUALITY OF SERVICE CLASSES AND A MILLION FLOWS.

Kenneth Y. Yun
Applied Micro Circuits

••••• To keep up with the explosive demand for bandwidth as well as to adhere to service-level agreements for a growing number of mature business applications on the Internet, network switches must be both faster and smarter. They must not simply terminate high-speed optical connections (OC-192 now and OC-768 in the near future). They also must switch a large number of connections from dense wave division multiplexing (DWDM) transport systems. They must provide guarantees on parameters such as bandwidth, latency, loss rate, and jitter that aren't supported by current best-effort switch architectures. Finally, they must provide a path to migrate to predominantly Internet protocol/multiprotocol label switching (IP/MPLS)-based networks without abandoning existing investments in the legacy network services such as time division multiplexing (TDM) and frame relay.

This article describes a terabit multiservice switch architecture designed to solve these problems. Because of its multiterabit switching speed, quality of service (QoS) support capability, multiprotocol capability including TDM, and scalability, the Cyclone switch architecture is directly applicable in the following areas and many more:

- multiterabit switching at the Internet core: terabit routers and carrier-class ATM or MPLS switches,
- aggregation for all optical networks;

- unified packet/circuit switching platform, and
- high-end enterprise applications.

Figure 1 shows a typical application of the Cyclone switch architecture. In this configuration, the switch core is physically separated from the rest of the switch and router. (This architecture also supports systems with integrated switch core and line termination cards.)

Cyclone switches use vertical cavity surface emitting laser, or VCSEL, arrays optimized for short-reach optical connections in this configuration. The current VCSEL technology can transmit up to 100 meters (extending to 500 m in the near future) at a 2.5-Gbps speed with low bit error rates. Therefore, Cyclone switches don't need to share the same rack space as the line cards containing PHY, framers, and network processors. In fact, Cyclone switches can use midrange switch routers as line cards.

Switch architecture overview

The Cyclone switch architecture is optimized for 32-port input and output buffered switches. Each port supports up to an 80-Gbps link bandwidth. Therefore, the aggregate bandwidth of a Cyclone switch is 5 terabits per second (2.5-Tbps ingress plus 2.5-Tbps egress). Each port consists of up to four channels, each of which supports up to 20 Gbps. Figure 2 shows a conceptual block diagram of the Cyclone switch architecture.

To prevent head of line (HOL) blocking problems, Cyclone switches use a virtual output queue (VOQ) for each output channel, for a total of 128 logical VOQs per input port. Both unicast and multicast transmissions are handled with a single mechanism, so that multicast packets incur no performance degradation.

The Cyclone switch architecture supports QoS at wire speed. In other words, Cyclone switches can provide advanced QoS features such as fair bandwidth allocation, low delay jitter, and priority queuing without sacrificing the switching speed. The Cyclone architecture accomplishes this by providing three levels of scheduling.

1. At the input side, the switch performs priority-based input sorting, using up to eight classes of service queues for each output channel. To select candidate packets for transmission (one for each destination), the switch uses either the deficit round-robin (DRR) algorithm¹ or the weighted round-robin (WRR) algorithm.
2. In its center, the switch uses a parallel arbitration algorithm for the maximal matching of packets at the heads of virtual output queues and their destination output channels.
3. At the output side, the switch implements either a programmable weighted fair queuing (WFQ)² or DRR algorithm for fair allocation of bandwidth.

In addition, the Cyclone architecture supports the provisioning of TDM service with an

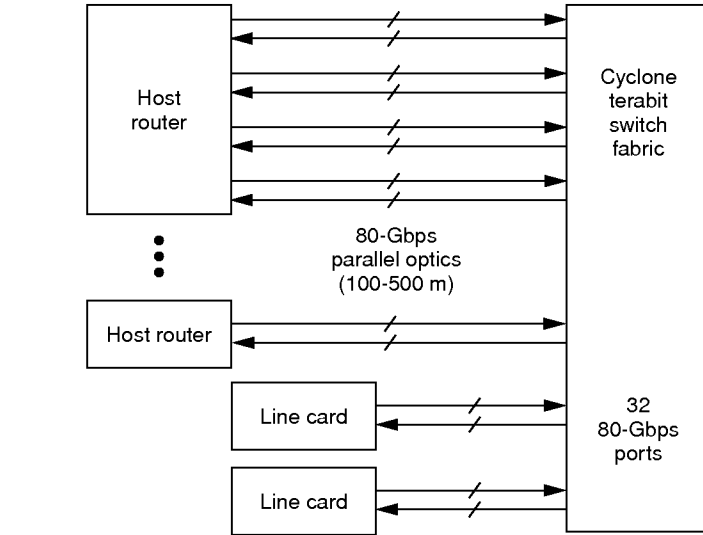


Figure 1. A typical application of the Cyclone switch architecture.

absolute guarantee of reserved bandwidth and zero delay jitter. Thus a Cyclone switch can be configured—a fraction of bandwidth is reserved for circuit switching and the remainder for packet switching—as a true multiservice switch.

Finally, the Cyclone switch architecture is cleanly scalable from 2.5 Tbps down to 40 Gbps without requiring design changes in the chip set.

Packet segmentation

Cyclone switches assume that packets are segmented into 64- or 80-byte fixed-size cells before entering the ingress of the switch. Each unicast cell contains a 6-byte header, including a special head of the cell indicator as the first

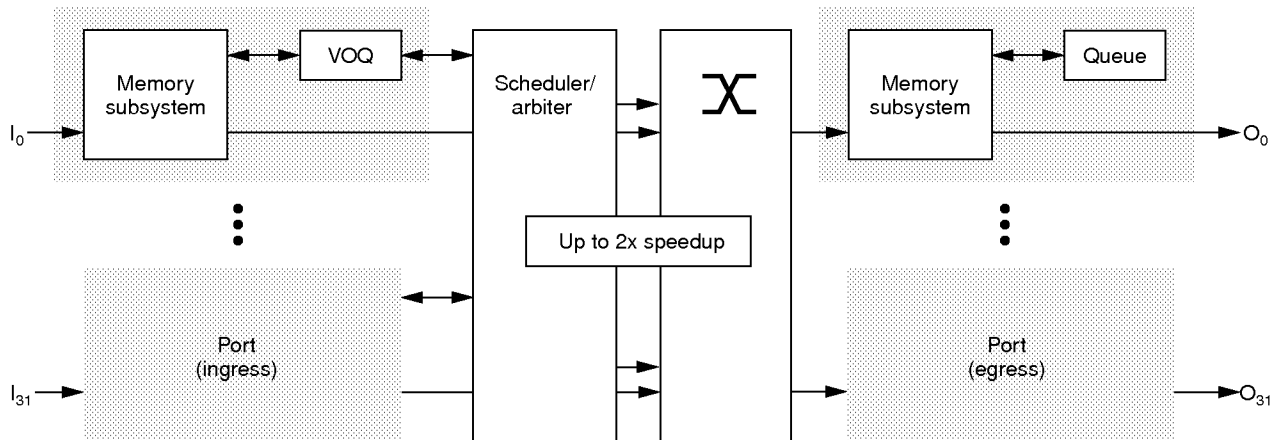


Figure 2. A conceptual block diagram of a 32 x 32 Cyclone switch architecture. VOQ = virtual output queue

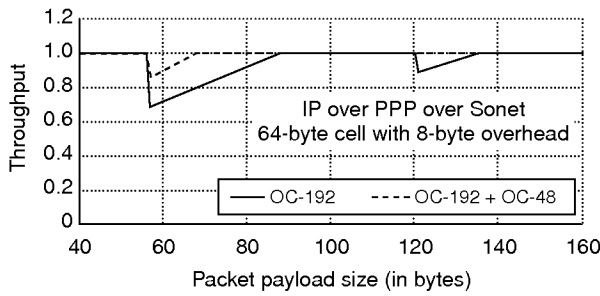


Figure 3. Throughput versus payload size (IP PPP over Sonet).

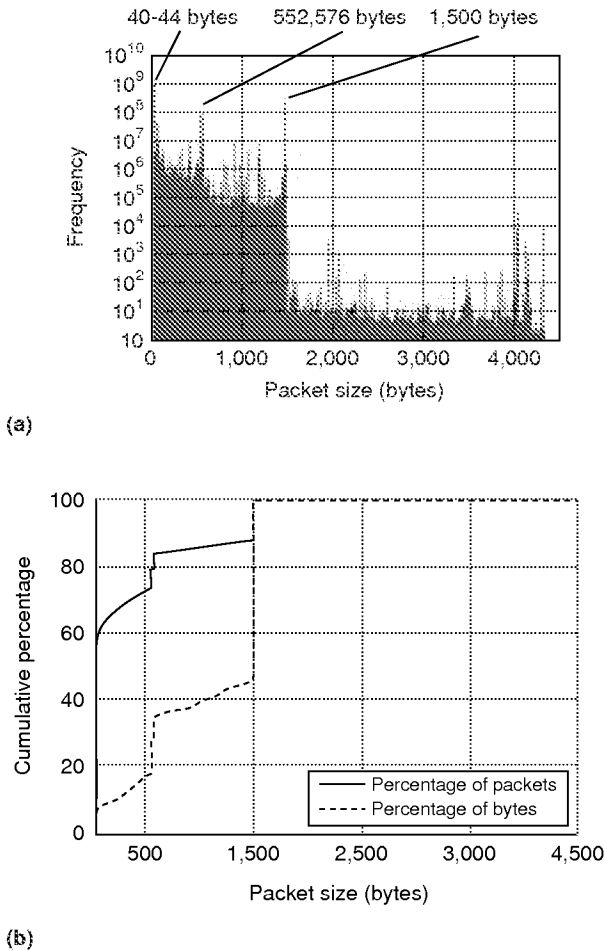


Figure 4. Internet packet distribution (MCI Study, 1998³): relative frequency of various packet size (a) and cumulative distribution (b).

byte (typically, a comma character if the incoming data are 8B10B-encoded), 56- or 72-byte payload, and a 2-byte vertical parity (or CRC) field. Thus the total cell overhead, is 8 bytes. Each multicast cell contains a 9-byte header, 53- or 69-byte payload, and a 2-byte parity field.

Although each channel of Cyclone switches can handle a 20-Gbps raw data rate, the actual payload throughput is lower due to the cell tax and the round-off overhead introduced by the cellularization of packets.

For example, each Sonet/SDH (synchronous digital hierarchy; the international version of Sonet) frame for STS-N (OC-N) consists of $N \times 90$ columns \times 9 rows = 810 N bytes, of which 3 columns are transport overhead bytes. A frame is transported every 125 microseconds. Therefore, $9 \times (90 - 3) \times N$ bytes (for example, 150,336 bytes for OC-192) are transported every 125 microseconds.

Furthermore, for packets transmitted over Sonet (IP over the point-to-point protocol, or PPP, over Sonet), each IP packet is associated with a 9-byte pause overhead. Hence, the actual payload size transported is $87/90 \times 810 N / (p + 9)$ packets or $783 N p / (p + 9)$ bytes per 125 microseconds. Here, p is the average packet size in bytes.

Assuming that all packets are unicast and the cell size is 64 bytes (56 byte payload plus 8-byte overhead), a channel of a Cyclone switch connected to line cards that terminate Sonet OC-N carrying IP packets over PPP, must be able to transport $783 N \times [64(p/56)] / (p + 9)$ bytes per 125 microseconds. This means that the switch must be sped up by a factor of $87/90 \times [64(p/56)] / (p + 9)$. For example, if $p = 40$, the speedup factor must be greater than 1.26.

Figure 3 shows the throughput that a Cyclone switch can sustain when each channel is connected to a Sonet OC-192 input stream carrying IP packets, or an OC-192 plus an OC-48, for a given IP packet payload size (assuming that every packet is of the same size). Clearly, for certain packet sizes, it's not possible to sustain 100% throughput. Furthermore, the higher the offered load, the more difficult it is to sustain the 100% throughput. However, as shown in Figure 4, the distribution of packet payload size dictates that the speedup factor should be optimized for 40- to 44-byte packets (TCP/IP acknowledge and control packets).

Port interface

A fully configured Cyclone switch contains 32 input/output port cards. Each port, as shown in Figure 5, can transmit and receive 8B10B-

encoded data at up to 100 Gbps (80-Gbps decoded data). Each port is organized as four channels. Thus each channel can transmit and receive 8B10B-encoded data at up to 25 Gbps.

Each channel consists of up to 8 to 10 physical serial links in each direction, and each link transports data at 2.5 Gbps. Cyclone switches configured with optical links use VCSEL arrays for transmitting and receiving signals on optical links.

Architectural details

The Cyclone switch can be configured as a unified packet/circuit-switching platform. Figures 6 and 7 illustrate the packet and TDM flows.

Packets entering the switch are stored in the input buffer and their pointers and lengths in the appropriate bins of the virtual output queues, according to their destination channels and class of service levels.

VOQs and arbiters negotiate and select the packets to be transmitted across the backplane. Selected packets are switched at the crossbars, according to the decisions made by the VOQs and the arbiters, and then transmitted to the egress side. Packets are written into the output buffer and their pointers into the appropriate output queues (labeled as EDFQ in Figure 6), where they are sorted, based on their priority levels, and selected for transmission.

TDM frames must be cellularized in the Cyclone cell format before entering the switch. Upon entrance, the frames are stored in the input buffer and their pointers in the TDM queue (separated from packet traffic). TDM frames are switched at the crossbars, as in the case of packets; however, the switching decisions are preprovisioned at the arbiters. TDM frames are then stored in the output buffer and their pointers in the TDM FIFO queue (segregated from

2.5-Gbps link (8B10B encoded)

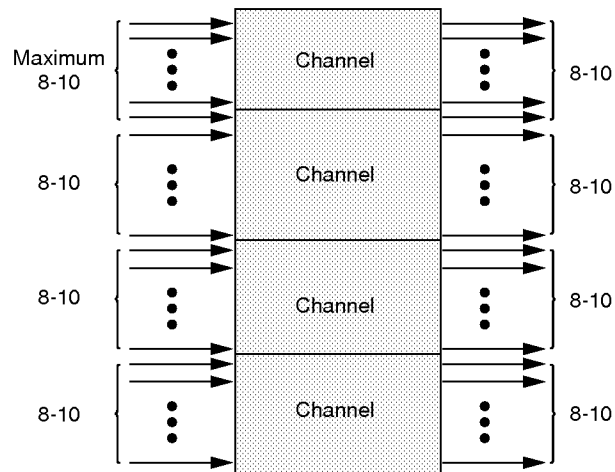


Figure 5. Port organization.

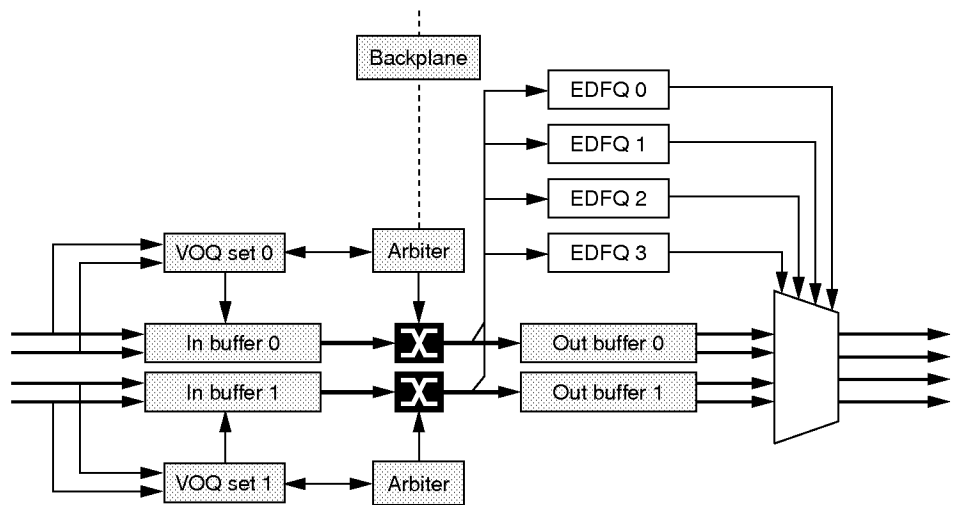


Figure 6. Packet flow assuming four channels per port. EDFQ = earliest deadline first queue

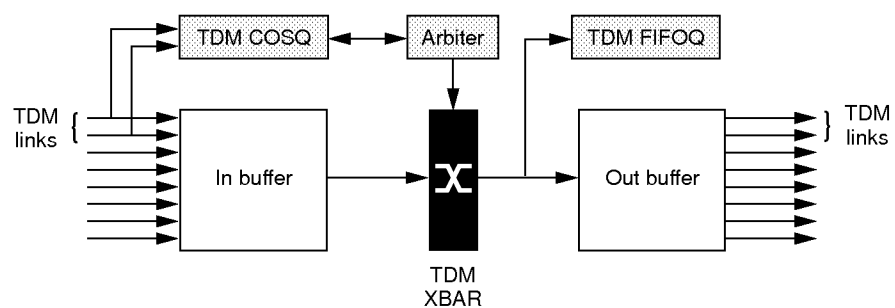


Figure 7. TDM flow. COSQ = class-of-service queue; FIFOQ = first-in, first-out queue

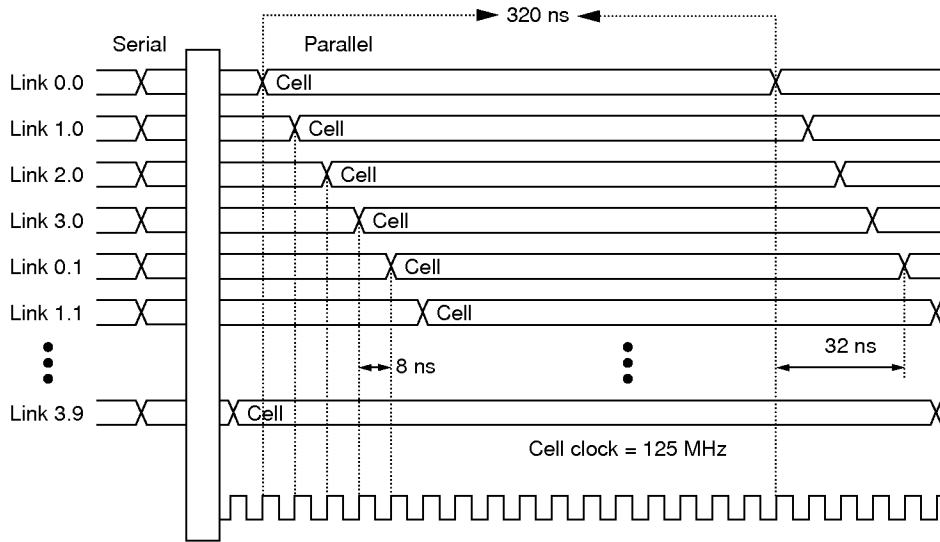


Figure 8. Input timing: cell size is 80 bytes; link speed is 2 Gbps; there are 10 links per channel and 4 channels per port.

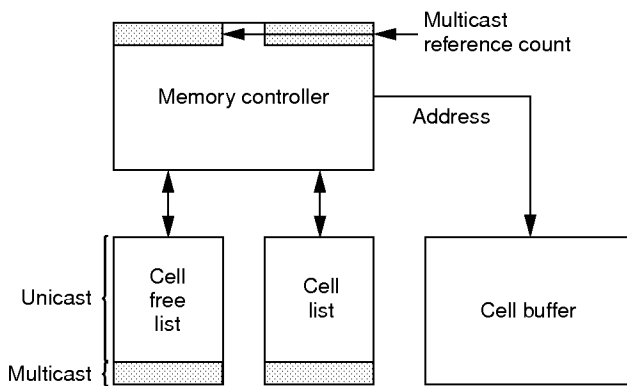


Figure 9. Memory controller.

the packet queues). They exit on the output links designated for TDM traffic.

Ingress

The ingress side of a Cyclone switch consists of a memory subsystem (which includes both the memory buffer and the controller) and virtual output queues.

Each packet in its entirety is carried on a link; that is, packets aren't "striped" across multiple links. Packets arriving at each port are assumed to be grouped by channels. Cyclone switches guarantee that packets belonging to the same flow (and arriving on the same channel) are transmitted in the order

they arrive. However, packets arriving on channel *i* are independent of packets arriving on channel *j*. In other words, the FIFO ordering of packets isn't maintained across the channels.

Arriving cells are parallelized in the manner depicted in Figure 8 before they are stored in the memory buffer one cell at a time every 8 ns. Cell parallelization is staggered by 8 ns so that the memory controller and the memory buffer process only one cell every 8 ns.

Since a cell time (the amount of time it takes to transport a cell) on a 2-Gbps link is 320 ns for cell sizes of 80 bytes, the maximum number

of 2-Gbps links per port (for a switch configured for an 80-byte cell size) is 40. If a port comprises more than one channel, the servicing of the channels is interleaved. An example is link 0 of channel 0 followed by link 0 of channels 1, 2, and 3 followed by link 1 of channels 0, 1, 2, and 3. This scheme has the effect of maximizing the distance between two adjacent links in the same channel so that the switch can better tolerate the arrival time jitter between two consecutive cells arriving on the same channel.

Memory controller. The memory controller, as shown in Figure 9, manages the memory buffer as well as the linked lists for packets and free cell slots.

- When a cell arrives, the memory controller assigns a cell pointer. If the cell is the head of a packet, it also assigns a packet ID, which is sent to the queue. The cell is then written into the memory buffer.
- When a packet is selected for transmission, cell pointers are retrieved using the packet ID. Packets are read out of the memory buffer one cell at a time, and cell pointers are recycled.

The memory controller also assists in the multicast management. When a multicast

packet arrives, it's stored in the input memory buffer, and its ID is copied to virtual output queues of all intended destinations simultaneously. After its packet ID is copied to the VOQs, the packet is copied to each destination, according to each VOQ's scheduling criterion. The packet remains in the input buffer until every intended destination receives a copy of the packet. The memory controller maintains the reference count of each multicast packet. Each time a destination receives a copy, the count is decremented. When the reference count of the packet reaches zero, its cell pointers are recycled: The packet is removed from the input memory buffer.

Backplane speedup. An input-blocking problem occurs when packets from multiple input ports destined for the same output port simultaneously appear at the heads of the corresponding input queues. If the output port can only receive one packet at a time, all but one packet are blocked. To minimize this problem, designers can configure the backplane bandwidth as twice the input bandwidth. That is, the Cyclone switch allows the backplane to be sped up by a factor of two. This speedup is accomplished by increasing the number of backplane links and by allowing up to twice as many cells to be read out of the memory buffer as written in. As shown in Figure 10, the backplane links are serviced in the same staggered fashion as the input links: links 0.0.0 and 1.0.0 followed by links 0.1.0 and 1.1.0, and so on.

Cyclone switches with four channels per port use two input memory banks per port to facilitate backplane speedup. (An important design criterion was to avoid using an expensive memory design. Each memory bank is implemented as a dual-port memory with an 8-ns cycle time.) Packets from input channels 0 and 2 are stored in bank 0, and packets from input channels 1 and 3 in bank 1. Both memory banks are accessed simultaneously. Half the backplane links are assigned to bank 0 and the other half to bank 1.

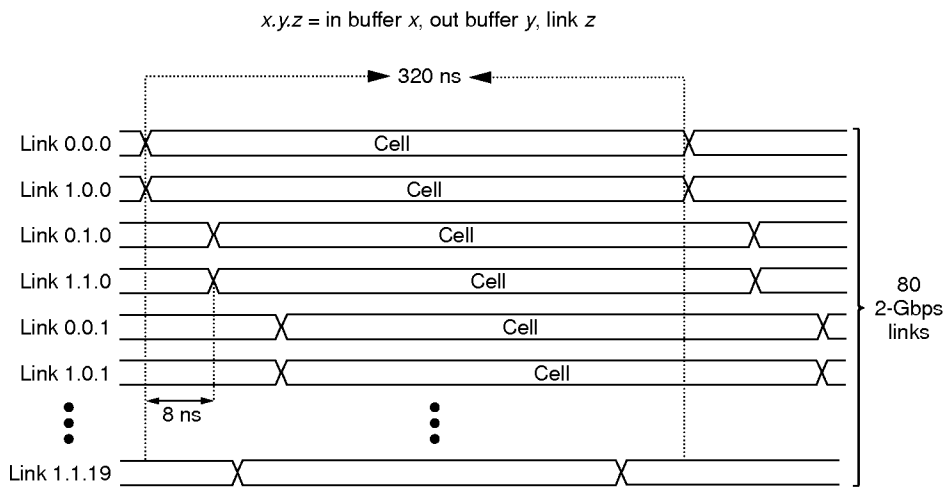


Figure 10. Backplane link timing. Link x.y.z means link number z on the switching plane associated with input buffer x and output buffer y.

Virtual output queue. The Cyclone switch uses VOQs to prevent head-of-line blocking problems.⁴ A separate VOQ is assigned to each output channel. Systems with 32 ports and 4 channels per port require 128 logical VOQs. Each VOQ entry consists of a pair (packet ID, packet length).

Recall that to support backplane speedup, two memory reads are required each cycle. To guarantee that two packets selected for transmission are from different banks, the Cyclone switch uses two sets of VOQs. One set handles packets stored in bank 0 (packets from input channels 0 and 2), and the other handles packets stored in bank 1 (packets from input channels 1 and 3).

Furthermore, each VOQ is organized as up to 8 class of service queues (COSQs) for service differentiation. Thus the maximum number of COSQs is 2,048 (2 input memory banks × 32 output ports × 4 channels per port × 8 COS levels).

Input scheduling. Each VOQ determines a candidate packet for transmission using one of two algorithms: DRR for variable-length packet traffic and weighted round robin (WRR)⁵ for fixed-length cell traffic. These algorithms select a candidate packet for each VOQ, from one of the (up to) eight COS queues that comprise a VOQ. Both DRR and WRR are system-configurable and can be modified to have up to four strict priorities.

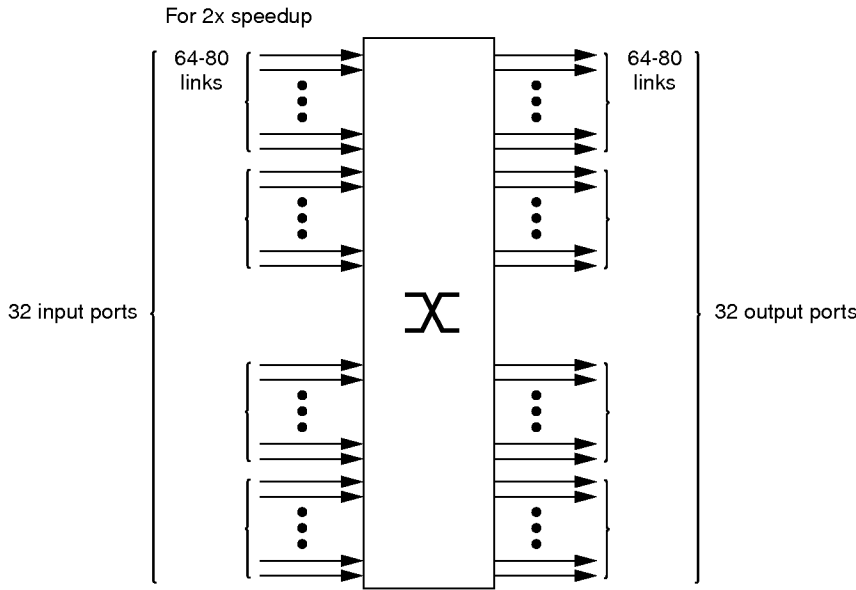


Figure 11. Backplane configuration.

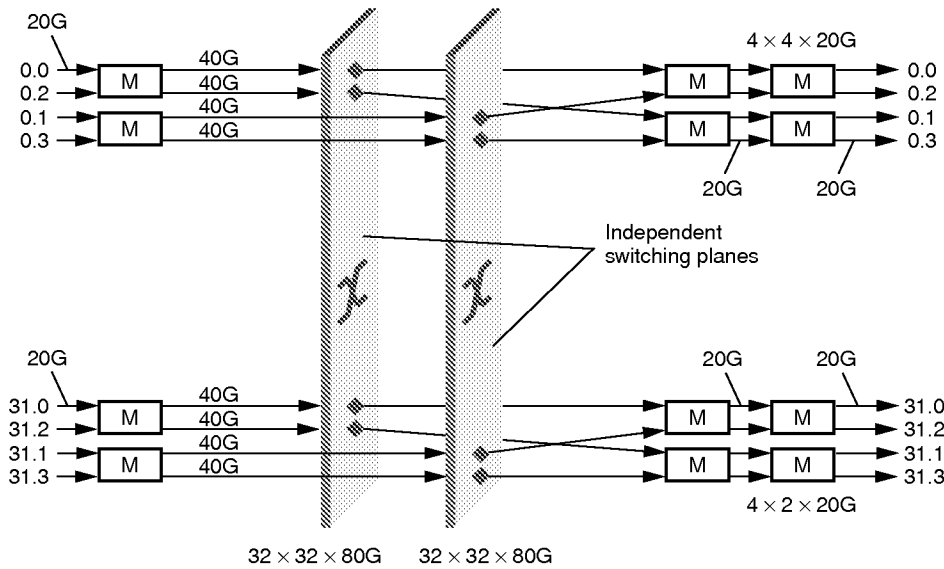


Figure 12. Channelized switching. $x.y$ = port x , channel y ; M = memory; G = Gbps

Only DRR is described here.

The DRR algorithm selects a candidate packet based on the following criterion. COS queues are visited in a round-robin fashion; only the backlogged queues are serviced. Each backlogged queue at its turn serves up to its credit, which is measured in cells and proportional to the weight assigned to each class. Any unused share is added to the credit for the next turn. For example, if the nominal credit is 8,

and the two packets at the head of queue contain 6 and 3 cells, only the first packet is served and the credit for the next turn becomes 10. If the queue is empty, its credit remains zero until it is backlogged again.

Backplane

With a speedup of two, the number of links from each input to the backplane is 80 for a cell size of 80 bytes and 64 for a 64-byte cell size, as depicted in Figure 11. Likewise, the number of links from the backplane to each output is 80 (64). As in the case of packet transmission on input links, each packet is carried on a link in its entirety—packets aren't striped across multiple links. The placement of cells on two adjacent pairs of links is staggered by 8 ns, as shown in Figure 10: Cells are placed on a pair of links every 8 ns. Likewise, the arrival of cells on two adjacent pairs of links from the backplane to an output port is staggered by 8 ns. Therefore, it takes 320 ns to service all 80 links (256 ns for 64 links), which is equal to one cell delay on a 2-Gbps link.

Parallel arbitration. As shown in Figure 12, two logical switching planes are used in Cyclone switches with four channels per port. One switching plane connects to one

input memory bank and the other plane to the other input memory bank. Each switching plane allows up to $2x$ backplane speedup; outgoing packets from each switching plane are sent to two output memory banks.

Each logical switching plane comprises multiple (up to eight) physical switching planes, each of which is controlled by an arbiter, as shown in Figure 13. Physical switching planes operate concurrently (with

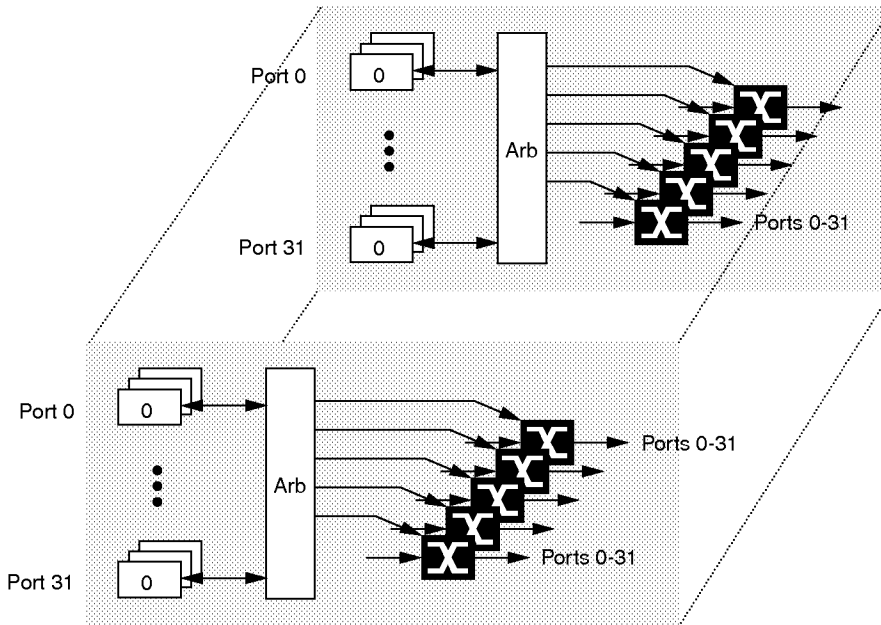


Figure 13. Parallel arbitration scheme.

8-ns staggering between adjacent planes) to allow multiple concurrent packet transmission from one input port to another. Note that no packet-ordering problem exists, even though each arbiter makes independent arbitration decisions. This is because each COSQ maintains a FIFO ordering of packets.

Each of the 80 sets of links (for switches with a cell size set to 80 bytes and a full 2x backplane speedup) is mapped to a crossbar. Up to five crossbars are assigned to each physical switching plane. An arbiter assigned to a physical switching plane makes arbitration (scheduling) decisions for all the crossbars in the plane. All the crossbars must be reconfigured in 320 ns. Since each arbiter is responsible for reconfiguring 5 crossbars in 320 ns sequentially (one crossbar at a time), arbitration decisions—each of which applies to a different crossbar—are made every 64 ns. Note that a new reconfiguration decision for each crossbar is made every 320 ns.

Hierarchical arbitration and VOQ/arbiter communication protocol. Each instance of the arbitration algorithm (for each switching plane) obtains the maximal number of matches (up to 32) between input ports and output ports. A type of maximal matching algorithm called SLIP was introduced by McKeown.⁴ However, our version distinctively differs from McKe-

own's in that arbitration decisions apply to packets, not just cells, and that multiple instances of the algorithm run on multiple switching planes (one for each plane).

Furthermore, if each port is configured with four channels, up to four packets can contend for each output port from each input port at any given time (one packet for each channel). The Cyclone switch uses a hierarchical arbitration algorithm to funnel the number of packets to at most one for each output port from each input port, as shown in Figure 14, next page.

Two communication links are used between each input port and an arbiter—one link each direction. The link is the same type used for data, but it's separate from data links: The control links are sideband. Note that arbitration decisions are made on a per-packet basis, not a per-cell basis. Once a decision is made on a link, the link is locked for the duration of the packet transmission on the link. The communication protocol follows:

1. VOQs from every input port make bids (and unlock the output links on which the packet transmissions have been completed).
2. The arbiter computes a match based on the maximal matching algorithm.
3. The arbiter accepts a bid for every un-

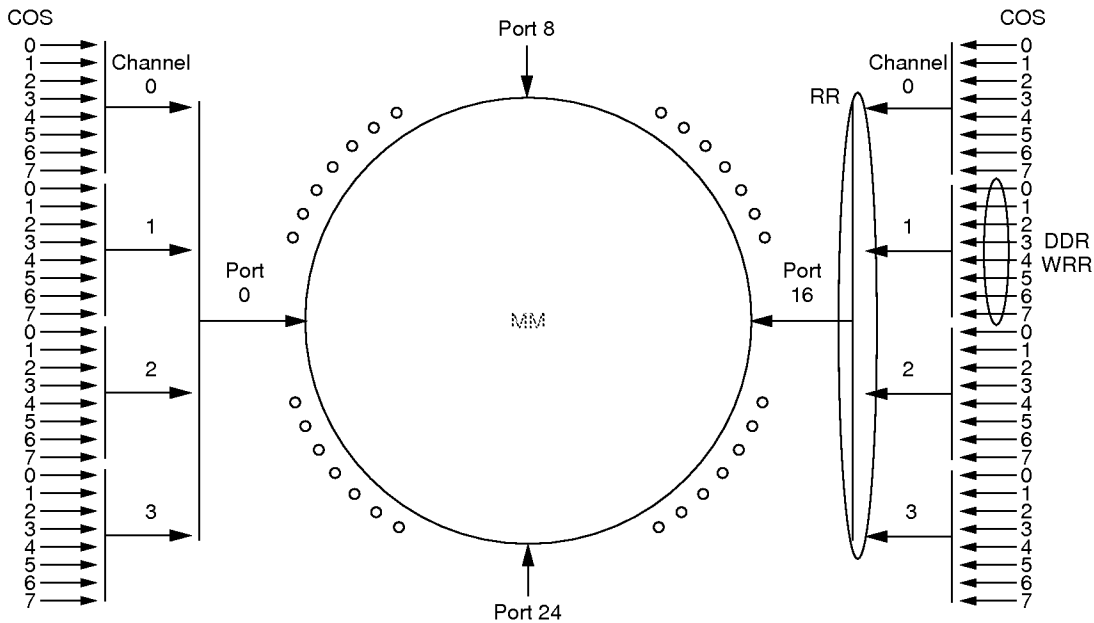


Figure 14. Hierarchical arbitration scheme. DRR = deficit round robin; WRR = weighted round robin; MM = maximal matching

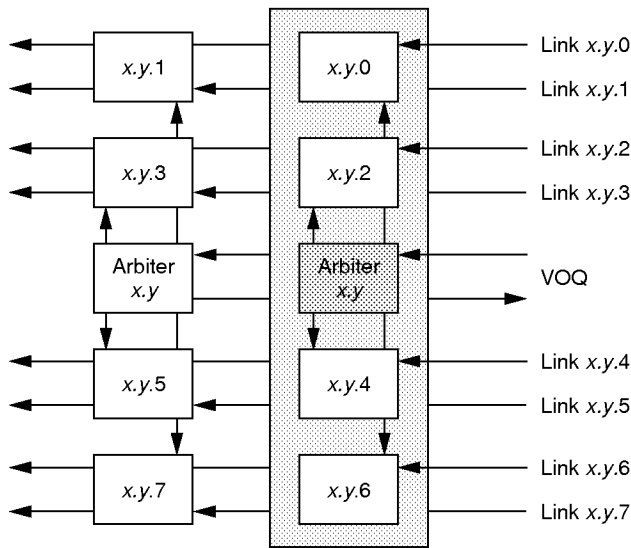


Figure 15. Backplane redundancy.

locked output port and locks the corresponding output link for the duration of packet transmission across the backplane.

A bid/match/accept cycle requires multi-stage pipeline processing, which means that new bids may be made for the next link before receiving the acceptance message for the prior bids. If back-to-back bids are made

from a COS queue that contains only one packet and the first bid is accepted, the second one is voided.

Backplane fault tolerance and redundancy. In Cyclone switches, each packet is carried entirely on one link (not striped across multiple links). Thus a hard failure on a link simply shuts down one link. Because the switch uses a large number of links (64 or 80) from each input port to the backplane, a link failure causes a very small degradation in performance.

Furthermore, the backplane speedup of two can be viewed as 1 + 1 backplane redundancy, as long as every arbitration channel is split into two cards, as depicted in Figure 15. Unlike conventional switches, Cyclone switches operate both primary and redundant channels. A failure on one side causes some degradation in performance (latency increase), not a catastrophic shutdown.

Egress

The memory subsystem on the egress side is identical to that on the ingress side, except that the egress memory buffer is designed to accommodate two writes and one read every 8 ns, instead of one write and two reads. As on the ingress side, Cyclone switches with four channels per port use two memory banks per

port on the egress side. The egress side supports one output queue per channel.

Output scheduling. Cyclone architecture supports two types of scheduling algorithms on the egress side: DRR/WRR for coarse-grain scheduling and weighted fair queuing (WFQ) for fine-grain scheduling.

If DRR/WRR is selected, the scheduler selects a packet for transmission from one of the eight possible COS queues that comprise a queue. (The DRR/WRR algorithm used on the egress side is the same as the one on the ingress side.)

If WFQ is selected, the WFQ scheduler assigns a service deadline for each packet based on the flow to which it belongs, and then sorts the packets by the deadline and selects the packet with the earliest deadline for transmission. In the Cyclone architecture, the flow is defined as a class of service supported on each low-speed output connection, for example, OC-3 or DS-3. Thirty-two thousand flows are supported per port (or 1 million flows total in 32-port-by-32 port configurations). The earliest-deadline-first queue (EDFQ) then sorts the packets by the deadline and selects the packet with the earliest deadline for transmission.

As shown in Figure 16, the Cyclone architecture uses a type of WFQ called self-clocked fair queuing (SCFQ).⁶ The deadlines for the packets destined for output channel x are computed as

$$F_j[k] = \max(F_j[k-1], F_x) + L[k]/P_j$$

Here, $F_j[k]$ is the deadline of packet k in flow j , F_x is the deadline of the last packet that has left channel x , $L[k]$ is the length of packet k , and P_j is the bandwidth allocated to flow j .

To avoid the ambiguity associated with ordering the deadlines with finite resolution—sometimes referred to as the time rollover problem⁷—the Cyclone architecture takes the following approach: No two deadlines of the packets stored in the queue (as well as the incoming) can be more than half the range apart. The Cyclone architecture assumes that the difference in deadlines of two consecutive packets in the highest speed/highest priority flow (for example, OC-192c at COS level 7) is nominally the length of the first packet (in the

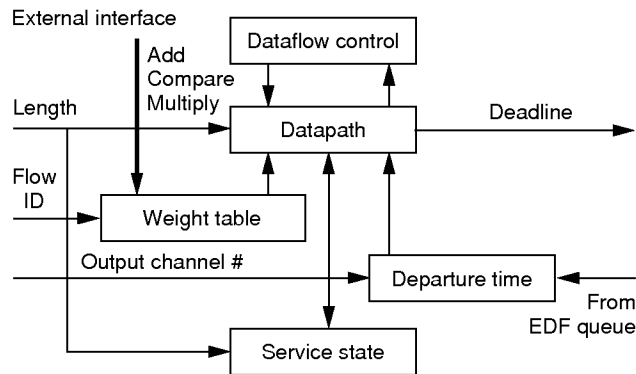


Figure 16. Self-clocked fair queuing scheduler.

number of cells). Thus the difference in deadlines of two consecutive packets in the lowest speed/lowest priority flow is $N \times M \times$ the length of the first packet. Here, N is the ratio of rates of the highest and the lowest speed connections allowed, and M is the ratio of bandwidths allocated for the highest and the lowest priority flows with the same speed. Hence, the maximum difference in the deadlines would be $L \times N \times M \times Q$. Here, L is the maximal transfer unit in cells, and Q is the number of queue entries. If $L = 256$, $N = 256$, $M = 32$, and $Q = 1,024$, the dynamic range of the deadline is 2^{31} . The Cyclone WFQ scheduler uses 32-bit deadlines, which constrains the deadline dynamic range to 2 billion time units.

EDFQ. This queue^{8,9} is a pipelined chain of discrete, locally interconnected stages (a systolic array). It sorts packet pointers by their deadlines so that the packet with the earliest deadline is selected for transmission every 8 ns. (The FIFO ordering is maintained for packets with the same deadline.)

The EDFQ receives two packet pointers and selects one packet for transmission every 8 ns. A new sorting process can be initiated every 4 ns since a new packet pointer arrives every 4 ns. To select the packet with the earliest deadline in a constant (and short) amount of time, the EDFQ always maintains the packet with the earliest deadline at the head of the queue.

Output timing. As in the case of packet transmission on input links, each packet in its entirety is carried on a link. The placement of cells on two adjacent links is staggered by 8 ns, because

the EDFQ can only transmit one packet every 8 ns. See Figure 17. In this example, four links are reserved for TDM and the rest for packets.

Flow control

The Cyclone switch architecture supports the flow control at the COS-level granularity. When a COS level is oversubscribed at an egress queue, the corresponding COS queues at the ingress side (for the output channel) are backpressured. This backpressure message must be broadcast to all of the corresponding COS queues on the ingress side, so the Cyclone switches use the arbiter to broadcast this message. Likewise, when a COS queue at the ingress side becomes nearly full, the external line card that feeds the COS queue is backpressured. This backpressuring can be localized to a COS level for a particular output destination channel.

Furthermore, the Cyclone switch architecture supports the separate flow control of each input channel. When packets from an input channel are oversubscribed at an output queue, the corresponding virtual output queue at the input channel is backpressured.

Simulation results

To quantify the performance of a typical configuration of the Cyclone switch architec-

ture, cycle-accurate behavioral simulator of a reference system was used with the following parameters:

- 32 full-duplex ports,
- 4 channels per port,
- 16-Gbps channel bandwidth;
- 64-byte cell size, 8-byte cell overhead, and
- a backplane speedup of 2.

Two types of packet length distribution were used: geometric distribution with a mean of 32 cells and MCI Internet distribution.³ The distribution over input and output channels was uniform. The offered load varied from 40% to 100%. A 70% load corresponds to the full OC-192 payload for both types of packet length distributions; 100% corresponds to approximately 1.4 times the OC-192 payload.

Figure 18 shows a latency comparison of the Cyclone switch configured as just described to a baseline switch configured as follows:

- 32 full-duplex ports,
- 1 channel per port; 10-Gbps channel bandwidth, and
- no backplane speedup (input buffering only).

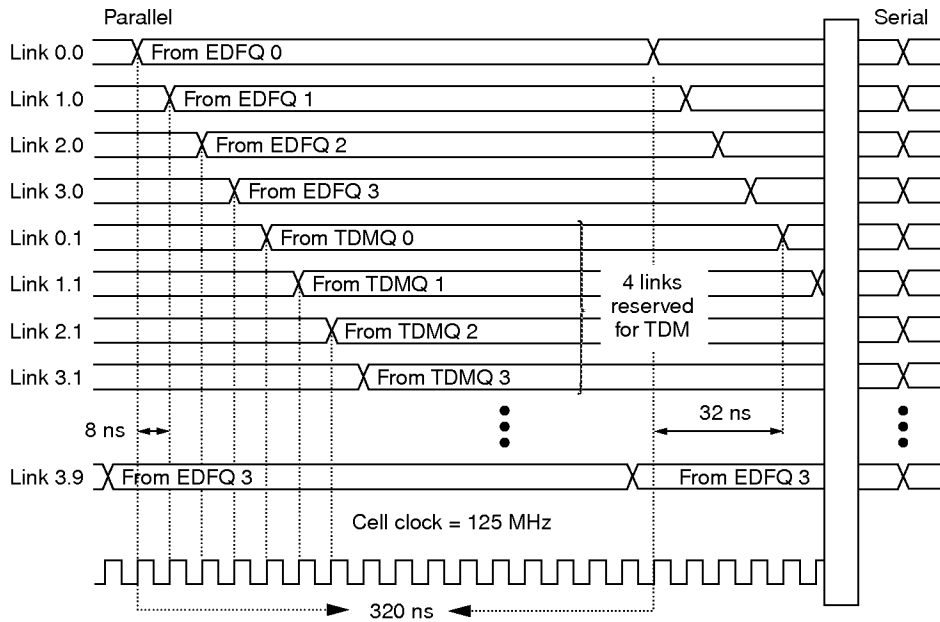


Figure 17. Output timing: cell size = 80 bytes; link speed = 2 Gbps; there are 10 links per channel and 4 channels per port.

The latency is measured in number of 8-ns clock cycles. As expected, the latency of the baseline switch increases exponentially as the offered load approaches 70% (100% of its channel bandwidth). However, the latency of the Cyclone switch remains flat up to a 90% offered load and increases to approximately 5,000 clock cycles at 100%.

Figure 19 shows peak buffer and queue occupancy measurement results for the MCI Internet distribution. Again, the distribution across input and output channels is uniform. Both the queue and buffer occupancies rise moderately exponentially as the offered load increases beyond

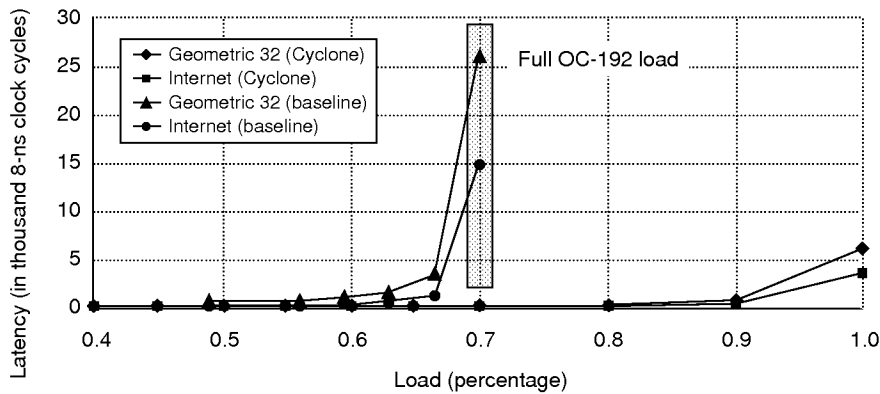


Figure 18. Latency versus offered load with 64-byte cells and 8-byte overhead. Baseline: 1x speedup; 32 10-Gbps channels. Cyclone: 1.6x speedup; 128 16-Gbps channels.

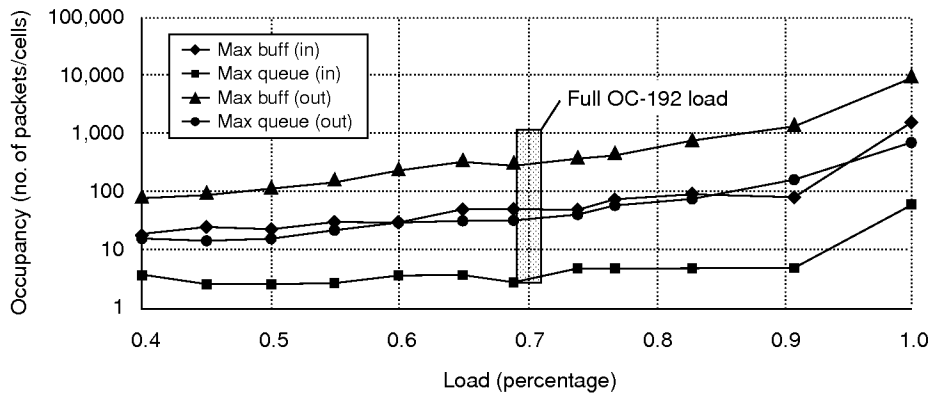


Figure 19. Queue/buffer occupancy versus offered load. Packet length = MCI Internet distribution; packet destination = uniform across 128 channels.

the full OC-192 load. Clearly, the ingress queue/buffer occupancy is one order of magnitude lower than the egress queue/buffer occupancy, which is expected with the significant backplane speedup.

The Cyclone switch architecture has several key attributes.

1. It's a true terabit switching platform; it can terminate and switch more than 128 OC-192c.
2. It supports multiple programmable scheduling algorithms, such as DRR, WRR, and WFQ, on the ingress side as well as on the egress side, which is necessary to assure end-to-end quality of service.
3. It's designed to support the TDM service naturally—for TDM traffic, the switch

guarantees reserved bandwidth and no delay jitter.

4. It's protocol-agnostic; it handles packet traffic and cell traffic equally well.
5. It's scalable from 40 Gbps to 2.5 Tbps without design changes.
6. It features robust fault tolerance and redundancy both at the link level as well as at the board level.

The Cyclone switch architecture has a potential to change the way future multiservice switches are designed because of its scalability, flexibility, and ability to handle multiple protocols equally well. MICRO

References

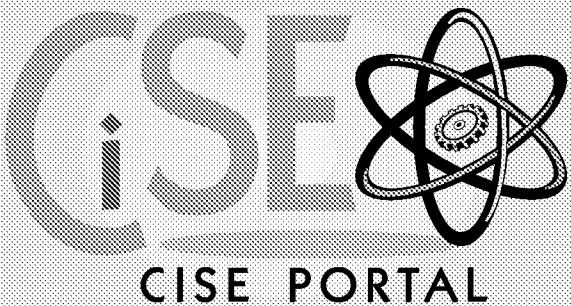
1. M. Shreedhar and G. Varghese, "Efficient Fair Queuing Using Deficit Round-Robin,"

- IEEE Trans. Networking*, vol. 4, no. 3, Jun. 1996, pp. 375-385.
2. A.K. Parekh and R.G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single Node Case," *IEEE/ACM Trans. Networking*, vol. 1, no. 3, Jun. 1993, pp. 344-357.
 3. K. Thompson, G.J. Miller, and R. Wilder, "Wide-Area Internet Traffic Patterns and Characteristics," *IEEE Network*, vol. 11, no. 6, Nov./Dec. 1997, pp. 10-23.
 4. N.W. McKeown, *Scheduling Algorithms for Input-Queued Cell Switches*, doctoral dissertation, EECS Dept., Univ. of California, Berkeley, 1995.
 5. M. Katavenis, S. Sidiropoulos, and S. Courcoubetis, "Weighted Round-Robin Cell Multiplexing in a General-Purpose ATM Switch Chip," *IEEE J. Selected Areas in Communications*, vol. 9, no. 8, Oct. 1991, pp. 1265-1279.
 6. S.J. Golestani, "A Self-Clocked Fair Queueing Scheme for Broadband Applications," *Proc. IEEE INFOCOM 94*, IEEE, Piscataway, N.J., 1994, pp. 636-646.
 7. H.J. Chao et al., "Design of a Generalized Priority Queue Manager for ATM Switches," *IEEE J. Selected Areas in Communications*, vol. 15, no. 5, Jun. 1997, pp. 867-880.
 8. K.W. James and K.Y. Yun, "A 40Gb/s Packet Switching Architecture with Fine-Grained Priorities," *Proc. Eighth Int'l Conf. on Computer Communications and Networks*, IEEE Press, Piscataway, N.J., 1999, pp. 148-153.
 9. K.W. James and K.Y. Yun, "Supporting Quality of Service in a Terabit Switch," *Proc. 19th IEEE Int'l Performance, Computing, and Communications Conf.*, IEEE Press, Piscataway, N.J., 2000, pp. 55-61.

Kenneth Yun is currently on leave from the University of California, San Diego to serve as a chief technical officer for Applied Micro Circuits Corporation. He has been an associate professor in the Department of Electrical and Computer Engineering at the University of California, San Diego and held design-engineering positions at TRW, Hitachi, Intel, AMD, and IBM. His current research interests include high-speed networking and mixed-timed design methodologies. Yun holds a PhD degree in electrical engineering from Stanford University and an SM degree in electrical engineering and computer science from MIT.

Direct comments about this article to Kenneth Yun, Chief Technical Officer, Applied Micro Circuits Corporation, 6310 Sequence Drive, San Diego, CA 92121; kyun@amcc.com.

COMPUTER.ORG/CISEPORTAL



A comprehensive, peer-reviewed resource for the scientific computing field.

Areas of expertise include

- Astronomy
- Chemistry
- Visualization
- Signal Processing
- Professional Resources

and more...